

Bertinho: Galician BERT Representations

Bertinho: Representaciones BERT para el gallego

David Vilares,¹ Marcos Garcia,² Carlos Gómez-Rodríguez¹

¹Universidade da Coruña, CITIC, Galicia, Spain

²CITIUS, Universidade de Santiago de Compostela, Galicia, Spain

david.vilares@udc.es, marcos.garcia.gonzalez@usc.gal, carlos.gomez@udc.es

Abstract: This paper presents a monolingual BERT model for Galician. We follow the recent trend that shows that it is feasible to build robust monolingual BERT models even for relatively low-resource languages, while performing better than the well-known official multilingual BERT (mBERT). More particularly, we release two monolingual Galician BERT models, built using 6 and 12 transformer layers, respectively; trained with limited resources (~45 million tokens on a single GPU of 24GB). We then provide an exhaustive evaluation on a number of tasks such as POS-tagging, dependency parsing and named entity recognition. For this purpose, all these tasks are cast in a pure sequence labeling setup in order to run BERT without the need to include any additional layers on top of it (we only use an output classification layer to map the contextualized representations into the predicted label). The experiments show that our models, especially the 12-layer one, outperform the results of mBERT in most tasks.

Keywords: BERT, Galician, embeddings, language modeling.

Resumen: Este artículo presenta un modelo BERT monolingüe para el gallego. Nos basamos en la tendencia actual que ha demostrado que es posible crear modelos BERT monolingües robustos incluso para aquellos idiomas para los que hay una relativa escasez de recursos, funcionando éstos mejor que el modelo BERT multilingüe oficial (mBERT). Concretamente, liberamos dos modelos monolingües para el gallego, creados con 6 y 12 capas de *transformers*, respectivamente, y entrenados con una limitada cantidad de recursos (~45 millones de palabras sobre una única GPU de 24GB.) Para evaluarlos realizamos un conjunto exhaustivo de experimentos en tareas como análisis morfosintáctico, análisis sintáctico de dependencias o reconocimiento de entidades. Para ello, abordamos estas tareas como etiquetado de secuencias, con el objetivo de ejecutar los modelos BERT sin la necesidad de incluir ninguna capa adicional (únicamente se añade la capa de salida encargada de transformar las representaciones contextualizadas en la etiqueta predicha). Los experimentos muestran que nuestros modelos, especialmente el de 12 capas, mejoran los resultados de mBERT en la mayor parte de las tareas.

Palabras clave: BERT, gallego, *embeddings*, modelado del lenguaje.

1 Introduction

Contextualized word representations (Dai and Le, 2015; Peters et al., 2018; Devlin et al., 2019) have largely improved the performance of many natural language processing (NLP) tasks, such as syntactic parsing (Kitaev and Klein, 2018), question answering (Salant and Berant, 2018) or natural language inference (Jiang and de Marnette, 2019), among many others. Contrary to static word embeddings (Mikolov et al., 2013b; Pennington, Socher, and Manning,

2014), where a given term is always represented by the same low-dimensional vector, contextualized approaches encode each word based on its context. Such process is normally learned by a neural network that optimizes a language modeling objective.

One of the most popularized and best performing models to generate contextualized representations is BERT (Devlin et al., 2019), a bidirectional language model based on transformers (Vaswani et al., 2017). BERT was initially released as a monolingual model

for English, with *large* and *base* variants, made of 24 and 12 transformer layers, respectively. In addition, a multilingual BERT version (mBERT) trained on the one hundred most popular languages on Wikipedia was also released. Although mBERT has become a very popular and easy-to-use tool to address multilingual NLP challenges (Pires, Schlinger, and Garrette, 2019; Karthikeyan et al., 2020), some authors have reported that its performance is not so robust as that of the corresponding monolingual models (Wu and Dredze, 2020; Vulić et al., 2020). In this line, previous work has showed that training a monolingual BERT is worth in terms of performance, in comparison to mBERT. Among others, this is the case for languages coming from different typologies, languages families and scripts, such as Finnish (Virtanen et al., 2019), Basque (Agerri et al., 2020), Spanish (Cañete et al., 2020), Greek (Koutsikakis et al., 2020) or Korean (Lee et al., 2020).

Taking the above into account, this paper contributes with the development of BERT models for Galician, a relatively low-resource language for which, to best of our knowledge, there is no contextualized (monolingual) model available. In this regard, we train two *Bertinho* models and test their performance on several tasks. Specifically, we assess the effect of the number of layers when using limited data (less than 45 million tokens). To do so, we train models with 6 and 12 layers using a single TESLA P40 24GB GPU and compare them against the official multilingual BERT on a number of downstream tasks, including POS-tagging, dependency parsing, and named entity recognition (NER). The experiments show that the monolingual models clearly outperform mBERT: even a small 6-layer model outperforms the official multilingual one in most scenarios, and the 12-layer one obtains the overall best results. We have submitted *Bertinho* to the HuggingFace Models Hub¹. Contemporaneously to this work, HuggingFace has released a Galician RoBERTa (Liu et al., 2019) model² based on the approach presented by Ortiz Suárez, Romary, and Sagot (2020).

¹<https://huggingface.co/dvilares/bertinho-gl-base-cased> and <https://huggingface.co/dvilares/bertinho-gl-small-cased>

²<https://huggingface.co/mrm8488/RoBERTinha>

Apart from this introduction, this paper is organized as follows: Section 2 introduces some related work regarding static and contextualized vector representations for NLP, discussing different approaches to train new monolingual models. Then, we describe the particularities of the Galician BERT-based models in Section 3. Finally, the experiments and results are presented and discussed in Section 4, while the conclusions are drawn in Section 5.

2 Related Work

The paradigm shift of NLP architectures produced by the rise of neural networks (Bengio et al., 2003; Collobert and Weston, 2008; Collobert et al., 2011) popularized the use of vector space models, following previous work in distributional semantics (Landauer and Dumais, 1997; McDonald and Ramscar, 2001). In this scenario, several highly efficient methods have been proposed to learn low-dimensional vector representation of words (i.e., word embeddings), such as *word2vec* (Mikolov et al., 2013a), *GloVe* (Pennington, Socher, and Manning, 2014), or *fastText* (Bojanowski et al., 2017). Since then, the use of pretrained embeddings to initialize the training of deep learning NLP models has become a standard procedure, due to the positive impact provided by the distributed representations in most downstream tasks (Schnabel et al., 2015).

One of the main drawbacks of these *static* word embeddings for NLP is that they represent all the senses of a given word in the same vector, thus making it difficult to deal with different linguistic phenomena such as polysemy or homonymy. In this regard, Peters et al. (2018) introduced ELMo, a model which obtains contextualized vector representations by means of an LSTM architecture, thus providing context-specific vectors for each token. This language model runs in a unidirectional fashion, in order not to trick itself when predicting the next word.

Following this idea, Devlin et al. (2019) presented BERT, a bidirectional language representation model based on the transformer architecture (Vaswani et al., 2017) and trained on a masked language model objective and on a next sentence prediction one, in order to consider both previous and upcoming context while still being a fair training objective. BERT has obtained state-of-

the-art results for many NLP tasks, and it is easy executable through freely available hubs, becoming a strong baseline in many recent work. In addition, the success achieved by BERT has promoted an extensive implementation of BERT-based models with several goals, such as optimizing its training method, e.g. RoBERTa, (Liu et al., 2019) or reducing the size and complexity of the model, e.g. DistilBERT (Sanh et al., 2019) or ALBERT (Lan et al., 2020)³. Moreover, it has attracted the interest of the research community regarding how its deep architecture encodes linguistic knowledge (Lin, Tan, and Frank, 2019; Vilares et al., 2020; Ettinger, 2020).

Besides the original English and Chinese models of BERT, the authors released a multilingual one (mBERT, with 12 layers) which produces inter-linguistic representations to some extent and gives good performance at zero-shot cross-lingual tasks (Pires, Schlinger, and Garrette, 2019). However, several studies have pointed that there are significant differences in performance among the languages covered by mBERT (Wu and Dredze, 2020), and that for some tasks even the best represented ones do not produce competitive results (Vulić et al., 2020). These findings suggest that, if possible, it is worth training monolingual models, especially for those languages which are poorly represented in the multilingual version of BERT.

In fact, monolingual versions of BERT have been trained for various languages using different methods, improving the results of mBERT. Finnish (Virtanen et al., 2019), Spanish (Cañete et al., 2020) and Greek (Koutsikakis et al., 2020) models were trained with about 3 billion tokens using the same parameters as the original BERT-Base (12 layers and 768 vector dimensions). For Portuguese, Souza, Nogueira, and Lotufo (2019) trained two models using about 2.7 billion tokens: one *large*, using the original English BERT-Large (with 24 layers and a vector size of 1024) for initialization, and one *base*, initialized from mBERT. A similar approach has been also adopted for Russian (Kuratov and Arkhipov, 2019), whose RuBERTa model used mBERT as the starting point, too. A comparison with a monolingual training with random initialization showed

that starting from the pre-trained mBERT reduces training time and allows for achieving better performance in various tasks. Finally, and more similar to our setting, the monolingual Basque model (Agerri et al., 2020) obtains state-of-the-art performance in different downstream tasks using only 225 million tokens for training a BERT-base model.

With the above in mind, this paper presents the work carried out to train two BERT models for Galician (one small, with only 6 layers, and one base, with 12), and evaluate them in downstream tasks such as POS-tagging, dependency parsing, and named entity recognition. Our approach can be seen as a low-resource scenario, as we only use one GPU and a small corpus of 42 million tokens for training.

3 Bertinho models

This section briefly introduces some ideas about Galician and describes the methodology that we have followed to train and evaluate the *Bertinho* models.

3.1 Galician

Galician is a romance language spoken by about 2.5 million people in the Spanish Autonomous Region of Galicia and adjacent territories (IGE, 2018). It belongs to the Western Ibero-Romance group, being evolved from the medieval Galician-Portuguese (Teyssier, 1987). Both philological and linguistic studies have traditionally classified Galician dialects as part of the same language as Portuguese (Lindley Cintra and Cunha, 1984; Freixeiro Mato, 2003), even though Galician has been standardized as an independent language since the 1970s, mainly through the use of a Spanish-based orthography (Samartim, 2012). In this regard, both Spanish and Portuguese NLP resources and tools have been used and adapted to analyse Galician data (Malvar et al., 2010; Garcia, Gómez-Rodríguez, and Alonso, 2018).

3.2 Training data for language modeling with BERT

For the pre-training phase, where BERT will be trained for language modeling in order to learn to generate robust contextualized word representations, we rely on a small corpus, extracted from the Galician version of the Wikipedia. More particularly, we used

³For which there is a monolingual Catalan model: <https://github.com/codegram/calbert>

the 2020-02-01 dump⁴ of the Galician version of the Wikipedia. To clean the data, we used `wikiextractor`⁵, which transforms the content of Wikipedia articles into raw text.⁶ We did not apply any further preprocessing steps, in order to do this training phase as self-supervised as possible. `Wikiextractor` divides the output into a number of text files of 1MB each. We selected the first 95% of these Wikipedia articles for the training set (with a total of 42 million tokens), and the remainder 5% for the dev set (2,5 million tokens), to keep track of the loss and perplexity at different training points and ensure a successful training. As an encyclopedic resource widely used by the NLP community, the resulting Wikipedia-based corpus is a well-structured and mostly clean dataset which does not contain as much noise as other crawled corpora (e.g. incomplete sentences, lines with no clear end, etc.).

Contrary to the original BERT release and some other monolingual trainings, we simply pre-train on the masked language objective and ignore the next sentence prediction one, since some recent BERT variants have shown that this second objective adds little or no benefit when it comes to fine-tune BERT-based models for downstream tasks (Liu et al., 2019), as we will be doing in this paper (see also Section 4).

3.3 Models

We now describe the procedure that we followed to pre-train our BERT models for language modeling, specifying the differences and similarities with respect to the training of other monolingual BERT models. We also will introduce the framework that we will use to fine-tune our models for downstream tasks.

3.3.1 BERT tokenizer and sub-word vocabulary

The BERT tokenizer splits the words into the so-called sub-word pieces (essentially n-grams of characters, given a word), where the least common words are simply represented by a generic unknown token, UNK. We follow the same setup as in the original English tok-

enizer, and define for both models a sub-word cased vocabulary of size 30,000. Such size was also set based both on the vocabularies used for BERT models for related linguistic varieties (e.g., Portuguese (Souza, Nogueira, and Lotufo, 2019) or Spanish (Cañete et al., 2020)), and on preliminary tests which suggested that this is a good trade-off between the size and the morphological correspondence of the sub-words.

Even though the tokenizer does not explicitly learn morphological information, better sub-words tend to correspond to morphological affixes, i.e., prefixes, suffixes or stems. In this regard, it is worth noting that other authors have used a larger vocabulary size to train BERT models in agglutinative languages, such as Basque (Agerri et al., 2020) or Finnish (Virtanen et al., 2019). Exploring an optimal vocabulary size for Galician falls out of the scope of this paper, but it might be an interesting future research line. Nevertheless, we show an example sentence tokenized by mBERT and by our model in Table 1. As it can be seen, mBERT splits the word *dixéronnos* ('they told us') into several sub-words, including 'dix', 'éro', 'nno', and 's'. Except for the first one ('dix') which corresponds to the frequent irregular root of the verb *dicir* ('tell'), the rest of the original token was split without taking into account morphological boundaries. However, the third sub-word identified by our model ('nos') is a correctly split clitic pronoun (masculine plural dative, 'to us'), while the second one ('éron') contains the thematic vowel and person morphemes. In addition, mBERT split the sub-word 'iño' from *camino* ('way' or 'path'), which in turn can be confused with the very frequent diminutive suffix 'iño' (e.g., the diminutive of *carro* -'car'- is *carrinho*), thus potentially involving inadequate representations of the whole word. Finally, the masculine possessive determiner *nosos* (with both plural possessor and possessed) was also split by mBERT (and not by our model, as it is a frequent word), but in this case it could be argued that the mBERT tokenization might not hurt the model, as the second sub-word 'os' corresponds to the masculine plural suffix, while the first one ('nos') could be analyzed as the morphological root (even though it also corresponds to a personal pronoun).

⁴Note that only the newest dumps are maintained over time <https://dumps.wikimedia.org/gwiki/>, but the differences should have a small effect in practice.

⁵<https://github.com/attardi/wikiextractor>

⁶In our work, we kept both the main texts and the headers, too.

Model	Tokenization
mBERT	Os nos ##os amigos dix ##éro ##nno ##s que o cam ##iño era este .
Ours	Os nosos amigos dix ##éron ##nos que o camiño era este .

Table 1: Tokenization of the sentence *Os nosos amigos dixéronnos que o camiño era este.* (‘Our friends told us that this was the way.’) by the original mBERT and our model. Following the same output representation as the BERT tokenizer, we use the symbol ## to specify a sub-word that is *not the first* sub-word of a split token.

3.3.2 Pre-training for language modeling

We have trained two models, varying the number of layers: (i) a BERT with 6 transformer layers (Bertinho_{SMALL}), and (ii) a BERT with 12 transformer layers (Bertinho_{BASE}), both trained on the Wikipedia corpus and using the tokenizer presented above. Each layer produces hidden representations of size 768, as in the original BERT paper for monolingual models. We use these models to explore the performance of each architecture in several downstream applications, thus being able to analyze the trade-off between the size and complexity of the models and their quality on extrinsic tasks.

With respect to the hyper-parameter settings, we mostly follow the standard pre-training configuration used in the original BERT paper. We use a learning rate of 1×10^{-4} with a linear weight decay of 0.01. We choose Adam as the network weight optimizer (Kingma and Ba, 2015) with $\epsilon = 1 \times 10^{-8}$. For the masked language objective and given an input sentence of length n , we mask randomly a 15% of the tokens. From those, 80% of them are replaced by the wildcard symbol [MASK], 10% of them are changed to a random word from the input vocabulary, and the remaining 10% are not modified.

With respect to the training process, in the original BERT model the authors train for 1M steps on sequences of 512 tokens and a batch size of 256. This strategy significantly increases the training time, as the self-attention included at each of the transformer layers in the BERT models runs in $\mathcal{O}(n^2)$, where n is the number of input tokens. Alternatively, authors such as Agerri et al. (2020) use a two-phase procedure using also a training batch size of 256. On the first phase, they train on sequences of length 128 during 900 000 steps. On the second phase, they

continue the training considering sequences of length 512 during 100 000 additional steps. However, even with this more modest training setup, the authors still had access to a few TPUs.

In our work, we stuck to a even lower computational resource setup, training the models on a single TESLA P40 GPU of 24GB. Following the standard approach, we first trained the 12-layer model, and consider a two-step training procedure (with the second phase being optional). For phase 1, we used a smaller training batch imposed by the hardware limitations. More particularly, we used training batches of size 96 considering sequences of 128 tokens. To counteract such smaller training batch, we trained instead the model during more steps, up to 2M. This phase 1 took 30 days to complete the training. Optionally, if the phase 2 was applied, we kept training the model from phase 1 using sequences of length 512. However, this required to limit our training batch size by a significant amount (more particularly, we could only fit 12 sequences in memory). We trained this second sequence for 1.4M additional steps (which took 4 extra days). Thus, for Bertinho_{BASE} we have an additional model trained with the two-phase strategy, that we will be referring as Bertinho_{BASE-2PH}. Figure 1 shows the evaluation perplexity obtained at different steps during phase 1 (shorter sequences) and phase 2 (longer sentences). However, we observed that this second phase was not useful to improve the results over the Bertinho_{BASE} model obtained after finishing phase 1 when it came to fine-tune models for downstream tasks (see also Section 4). For the 6-layers model, Bertinho_{SMALL}, we decided to apply only the phase 1 of the training procedure, but with batches of size 128 instead and training for 1.5M steps (taking about 22 days to be completed). Figure 2 shows the perplexity of the SMALL model during evaluation.

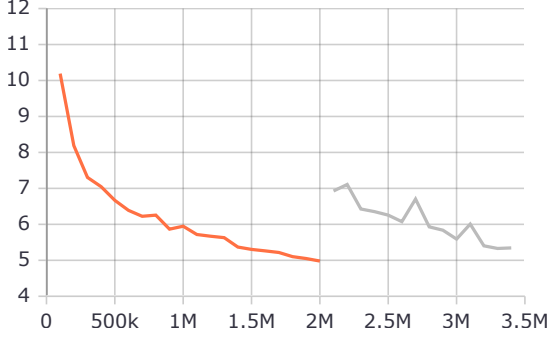


Figure 1: Eval perplexity for Bertinho_{BASE} (12 layers). The eval perplexity over sequences of 128 tokens during the phase 1 of the training is showed in red (left plot). The one over sequences of 512 tokens during the second phase is showed in gray (right plot).

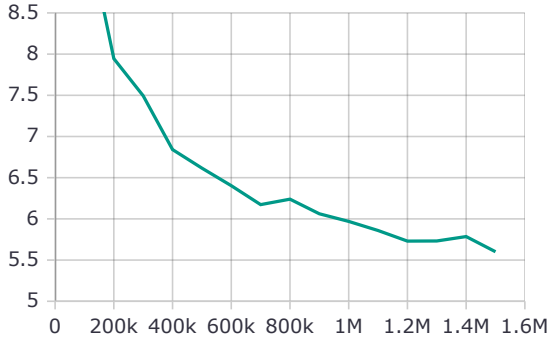


Figure 2: Eval perplexity for Bertinho_{SMALL}, trained only over sequence of 128 tokens.

All the models have been trained using the *transformers*⁷ library provided by the Hugging Face team (Wolf et al., 2019) using, as mentioned, a single 24 GB TESLA P40 GPU.

3.3.3 Framework

Given an input sentence $\vec{w}=[w_1, w_2, \dots, w_n]$, we propose to fine-tune the BERT pretrained models for language modeling to a number of tasks involving lexical, syntactic and semantic processing. We will cast these downstream tasks in a sequence labeling setup, i.e, we will learn a mapping $\phi: W^n \rightarrow L^n$ where L will represent the sequence of output labels. The tasks that we will study are POS-tagging, dependency parsing and named entity recognition, described in detail in Section 4.

That said, let BERT(\vec{w}) be a pre-trained BERT model that maps a one-input vector to a sequence of contextualized representations, \vec{h} , we simply map each h_i to an out-

put label y_i through a 1-layer feed-forward network using a softmax ($P(y_i = j|\vec{h}_i) = \text{softmax}(\vec{W} \cdot \vec{h}_i + \vec{b}) = \frac{e^{\vec{W}_j \cdot \vec{h}_i}}{\sum_k e^{\vec{W}_k \cdot \vec{h}_i}}$) to obtain a probability distribution over the set of output labels for each word, w_i . In our work, all labels are predicted in an atomic way, that is with single-task learning.⁸ In all cases, the loss, \mathcal{L} , is optimized based on categorical cross-entropy loss ($\mathcal{L} = -\sum \log(P(y_i|\vec{h}_i))$), and back-propagated through the whole network, i.e. we also fine-tune the BERT weights during the training for the downstream tasks.

4 Experiments and results

To evaluate the usefulness and robustness of the *Bertinho* models when it comes to fine-tuning the model for downstream tasks, we will consider three different problems: (i) part-of-speech tagging, (ii) dependency parsing, and (iii) named entity recognition. We will rely on existing datasets for Galician, which we now briefly review, indicating as well for which tasks we will use them:

- **CTAG corpus:** The *Corpus Técnico do Galego* (Galician Technical Corpus) is a linguistic corpus developed by the Seminario de Lingüística Informática of the University of Vigo, and it is composed of texts from a variety of technical domains (e.g., legal and scientific), totaling almost 18 million words⁹. A subset of the CTG corpus, the CTAG (Corpus Técnico Anotado do Galego, ‘Galician Technical Annotated Corpus’)¹⁰, was manually reviewed (Guinovart and Fernández, 2009; Agerri et al., 2018), therefore producing gold-standard resources for POS-tagging and lemmatization (with more than 2 million tokens (TALG, 2016)) and NER (with about 200k tokens (TALG, 2018)).

Tasks and observations: We will use these resources to evaluate the performance of the BERT-based models on POS-tagging and named entity recognition.

⁸Especially for more complex tasks, such as dependency parsing as sequence labeling, other authors (Strzyz, Vilares, and Gómez-Rodríguez, 2019) have decomposed the task into predicting partial labels using a multi-task learning setup.

⁹<http://sli.uvigo.es/CTG/>

¹⁰<http://sli.uvigo.es/CTAG/>

⁷<https://github.com/huggingface/transformers>

- **CTG-UD:** The CTG-UD (Guinovart, 2017) is a treebank based on a subpart of the CTG corpus which has been automatically parsed and adapted to Universal Dependencies¹¹ (Nivre et al., 2020).

Tasks and observations: CTG-UD will be used to evaluate both POS-tagging and dependency parsing. For POS-tagging, we will consider both universal and language-specific part-of-speech tags (based on the fine-grained tagset of FreeLing (Padr , 2011; Garcia and Gamallo, 2010)) as separate tasks. For dependency parsing, and to cast it as a sequence labeling task, we are relying on the encodings proposed by Strzyz, Vilares, and G mez-Rodr guez (2019), which defined different ways to map a sequence of input words of length n into a sequence of syntactic labels, also of length n , that can be decoded into full dependency trees. More particularly we will consider their bracketing encoding. Since it is outside of the scope of this work, we leave the details of the encodings for the reader, which can be found in the referenced work, and we simply use their encoding and decoding functions as a black box.

- **Galician-TreeGal:** This treebank is a subset of the XIADA corpus (Rojo et al., 2019) that has been annotated following the guidelines of the Universal Dependencies initiative. Galician-TreeGal has around 25k tokens, and its manually revised annotation includes lemmas, POS-tags, morphological features and dependency labels (Garcia, G mez-Rodr guez, and Alonso, 2016; Garcia, G mez-Rodr guez, and Alonso, 2018).

Tasks and observations: As in the case of the CTG-UD, we will use this dataset for universal and language-dependent POS-tagging (using the morphologically-rich XIADA tagset)¹², and also for UD dependency parsing. It is worth mentioning that due to the small size of the Galician TreeGal treebank, the evaluation can be considered as few-shot learning.

All the experiments have been carried

¹¹<https://universaldependencies.org/>

¹²<http://corpus.cirp.es/xiada/etiquetario/taboa>

out by fine-tuning the BERT-based models (Bertinho_{SMALL} and Bertinho_{BASE}) for each specific task using the splits for training and development of each corpus¹³ with the referred *transformers* library. As a baseline, we use the official mBERT released by Google (BERT-Base Multilingual Cased, with 12 layers), which includes Galician among its 104 covered languages.

4.1 Results

We first show the POS-tagging and NER results on the CTAG corpus, followed by the analyses on the Universal Dependencies treebanks (CTG-UD and Galician-TreeGal).

4.1.1 CTAG

Table 2 shows the POS-tagging and NER results on the CTAG corpus. The POS tagset contains 178 tags, and the NER labeling was approached as a BIO classification of four classes (person, location, organization, and miscellaneous) totaling 9 tags (two $-B$ and $-I$ for each class, and O for the tokens outside the named entities).

On POS-tagging, the best results are obtained by Bertinho_{BASE} (two-phase and single-phase, respectively), followed by Bertinho_{SMALL}, all of them surpassing mBERT in this task. Interestingly, even the 6-layer model obtained better results (2.32%) than the multilingual one, with 12 transformer layers. However, it is worth noting that mBERT outperformed all the monolingual models on named entity recognition, both in precision and recall, suggesting that for this task multilingual information may improve the performance. These and other results are discussed on Section 4.2.

4.1.2 Universal Dependencies

With respect to the Universal Dependencies treebanks, the results on the CTG-UD are shown on Table 3, including POS-tagging (on UD POS-tags and using a fine-grained tagset) and dependency parsing (both LAS and UAS values).

In this case, and also in further experiments, Bertinho_{BASE} obtained the best results in both tasks and settings, followed by

¹³Since Galician-TreeGal does not contain a development set, we have splitted the train data into train/dev with a 90%/10% ratio. Similarly, we have used 150k tokens from the training set of the CTAG for development of the POS-taggers.

Model	CTAG			
	POS	NER		
	Acc	P	R	F1
mBERT	93.84	83.53	85.60	84.55
Bertinho _{SMALL}	96.16	78.40	82.68	80.48
Bertinho _{BASE}	96.40	80.49	82.74	81.60
Bertinho _{BASE-2PH}	96.23	80.89	84.33	82.57

Table 2: POS-tagging and NER (precision, recall, and f-score) results on the CTAG corpus. The POS tagset corresponds to the fine-grained tags of FreeLing (see Section 3.2).

Model	UPOS	FPOS	LAS	UAS
mBERT	95.41	91.91	76.48	80.80
Bertinho _{SMALL}	96.42	94.56	77.59	81.55
Bertinho _{BASE}	96.56	94.60	78.14	81.88
Bertinho _{BASE-2PH}	96.43	94.50	77.70	81.65

Table 3: POS-tagging accuracies and dependency parsing results on the CTG-UD treebank. POS includes universal and language-dependent tagsets, while parsing is evaluated using LAS and UAS.

Bertinho_{BASE-2PH} and by Bertinho_{SMALL} (almost tied) and mBERT. As expected, the performance on POS-tagging is higher when using the UPOS tagset (with 16 elements) than with the FPOS one (194 tags). On dependency parsing, all the models follow the same mentioned tendency, with Bertinho_{BASE} achieving the best results: 78.14 and 81.88 (LAS and UAS, respectively).

When moving to the Galician-TreeGal treebank (Table 4), we see again that the best performance is achieved by Bertinho_{BASE} (single and two-phases, respectively), followed by the _{SMALL} variant, and finally by mBERT. It is worth recalling that this treebank is significantly smaller than the other datasets, therefore the model weights have more influence as the fine-tuning process is shorter. In this regard, the results of all models are lower than those obtained on the CTG-UD dataset, achieving POS-tagging accuracies of up to 96.61% (UPOS, with 16 tags) and 92.70% (FPOS, containing a tagset of 237 elements). On dependency parsing, the best results were of 75.26% and 80.27% on LAS and UAS, respectively.

Significance tests: We applied significance tests for POS-tagging and parsing, to determine whether the proposed monolingual models are actually different than mBERT. For POS-tagging, we applied a t-

test that compares the accuracies per sentence obtained by mBERT and each of the monolingual models. All models are significantly different (with $p < 0.01$), except Bertinho_{BASE-2PH} on the Galician-TreeGal (UPOS). For parsing, as in Vilares and Gómez-Rodríguez (2018), we used instead the Bikel’s randomized parsing evaluation comparator, a stratified shuffling significance test. The null hypothesis is that the outputs produced by mBERT and any of the monolingual models are produced by similar models and so the scores are equally likely. To refute it, it first measures the difference obtained for a metric by the two models. Then, it shuffles scores of individual sentences between the two models and recalculates the metrics, checking if the difference is less than the original one, which would be an indicator that the outputs generated by the models are significantly different. All models are significantly different (with $p < 0.01$).

4.2 Discussion

There is a clear tendency with respect to the performance of the different models on the downstream tasks. Except for NER (discussed below), Bertinho_{BASE} (12 layers) consistently obtains the best results, following by Bertinho_{SMALL} (6 layers) and the official mBERT release (also with 12 layers).

Model	UPOS	FPOS	LAS	UAS
mBERT	94.27	87.67	71.54	77.67
Bertinho _{SMALL}	96.38	92.05	73.23	78.71
Bertinho _{BASE}	96.61	92.70	75.26	80.27
Bertinho _{BASE-2PH}	96.46	92.69	74.41	79.64

Table 4: POS-tagging accuracies and dependency parsing results on the Galician-TreeGal tree-bank. POS includes universal and language-dependent tagsets, while parsing is evaluated using LAS and UAS.

Apart from the models themselves and from task-specific properties, there are two parameters that seem to play an important role when fine-tuning for a particular downstream task: the size of the tagset and the amount of training data. In this regard, the gain obtained by the best models (when compared to those with lower results) is higher when using large tagsets and smaller datasets, thus suggesting that they encode better information which in turn has a stronger impact on the final performance.

We have assessed this finding by observing the differences between the results of our best model (Bertinho_{BASE}) with those of mBERT: With respect to the size of the tagset, we compared the results on UPOS and FPOS on both Universal Dependencies corpora. The difference between Bertinho_{BASE} and mBERT on UPOS tagging (with 16 tags) was of 1.15 and 2.34 on CTG-UD and Galician-TreeGal, respectively. However, when using FPOS (with 194 and 237 different tags on the mentioned corpora), the gain achieved by Bertinho_{BASE} increased to 2.69 (CTG-UD) and 5.03 (Galician-TreeGal).

Besides the divergences between the UPOS and FPOS scenarios, these results also indicate that the differences are noticeably larger on the Galician-TreeGal than on the CTG-UD dataset. As the former tree-bank has less than 14k tokens for training (while CTG-UD has $\approx 80k$), the differences seem to be mainly caused by the disparity on the amount of training data. This tendency is also displayed on dependency parsing, where the LAS/UAS differences between Bertinho_{BASE} and mBERT are of 1.66/1.08 on CTG-UD and of 3.72/2.60 on Galician-TreeGal. Finally, this tendency does not hold when comparing the POS-tagging results on both CTAG and CTG datasets (2.56 *versus* 2.69 on the CTAG and CTG-UD, respec-

tively). Additionally, in this case the UD variant is 10 times smaller and has a larger dataset (194 *versus* 178), and consequently the results are lower than in the non-UD corpus (94.60 *versus* 96.40).

The mentioned tendency is not followed in the NER results, where the multilingual model achieved impressive performance (surpassing the best results published by Agerri et al. (2018)). Bertinho_{BASE} beats the SMALL variant by 1.12 points, but mBERT overcomes the best monolingual model (Bertinho_{BASE-2PH}) by 1.98 points. At first glance, we could hypothesize that the multilingual model performs better at NER as *enamel* named entities (locations, people, organizations, and miscellaneous entities) are represented interlinguistically, so that the model takes advantage of information from various languages. Although this may affect the results in some way, a careful analysis of the output of mBERT and Bertinho_{BASE} has shown that most errors of the monolingual model came from variation regarding upper and lowercase. Thus, in expressions such as “Especies más afectadas polo *Plano de Selado*” (‘Species most affected by the Sealing Plan’), “O *Código Civil* actual” (‘The current Civil Code’), or “As *Illas do Sur*” (‘The Southern Islands’), our model classified the expressions in italic as miscellaneous (the first two) and location (the last one) entities, but they are not labeled in the gold standard dataset. About the identification of person entities, Bertinho_{BASE} failed in some complex nouns including prepositions (e.g., “Bernardo Barreiro de Vázquez Varela” labeled as two entities –separated by ‘de’– instead of one), while some organizations named with common nouns were not identified by the NER system (e.g., “Cachoeira”, ‘waterfall’).¹⁴ Therefore, it could

¹⁴Some other errors of Bertinho_{BASE} were due

be interesting to analyze different techniques to deal with these issues in further work.

Finally, it is worth noting that the Bertinho_{BASE} variant trained with a two-phase strategy (Bertinho_{BASE-2PH}) consistently obtained worse performance than the single-phase one, except in the named entity recognition scenario. Even though more investigation is needed, these results indicate that in our case, and contrary to previous related work, a two-phase training procedure (training with longer sentences during the second phase) was not beneficial, hurting the performance of the model in the downstream tasks. We hypothesize this might be partially due to our limited hardware resources, that imposed us a very small training batch during the second phase of training.

In sum, we have trained and evaluated two BERT models for Galician that obtain better results than the official multilingual model in different downstream tasks. Interestingly, we provide both a 12-layer model (with greater performance), and a 6-layer one which obtains competitive results with a computationally less expensive architecture.

5 Conclusion

We have trained two monolingual BERT models for Galician (dubbed *Bertinho*), for which we have followed a low-resource approach with less than 45 million tokens of training data in a single GPU. Both models have been evaluated on several downstream tasks, namely on dependency parsing, NER, and POS-tagging with different tagsets. Moreover, we have shown how a dedicated tokenizer improves the morphological segmentation of Galician words.

Our best model (Bertinho_{BASE}, with 12 layers) outperforms the official BERT multilingual model (mBERT) in most downstream tasks and settings, and even the small one (Bertinho_{SMALL} with 6 layers) achieves better results than mBERT on the same tasks. However, it is worth noting that mBERT has worked better than the monolingual models on NER. Finally, our experiments have also shown that a two-phase training procedure for language modeling (with more learning steps and training with longer sequence at the end) consistently hurts the performance

of the 12-layer model on most scenarios in our setup. We believe this might be due to our limited hardware resources, that forced us to use a small training batch when pre-training with very long sequences.

In further work we plan to carry out a deeper analysis of the NER results, and also to compare the different layers of the BERT models with static word embeddings such as *word2vec* or GloVe. Furthermore, we aim to extend our models using larger corpora. *Bertinho* has been trained on less data than other BERT models for related languages such as BETO (Cañete et al., 2020). We believe collecting more data for other sources such as CommonCrawl could improve the performance. However it is also fair to state that there are also studies that point in the opposite direction. For instance, the results of Raffel et al. (2020) indicate that smaller clean datasets are better than large noisy corpora, so that it could be interesting to assess to what extent our results (obtained with a small dataset) can be improved with new data crawled from the web (Agerri et al., 2018; Wenzek et al., 2020).

Finally, it is important to recall that the work performed in this study contributes to the NLP community with the release of two freely available *Bertinho* models for Galician.

Acknowledgements

This work has received funding from the European Research Council (ERC), which has funded this research under the European Union’s Horizon 2020 research and innovation programme (FASTPARSE, grant agreement No 714150), from MINECO (ANSWER-ASAP, TIN2017-85160-C2-1-R), from Xunta de Galicia (ED431C 2020/11), from Centro de Investigación de Galicia ‘CITIC’, funded by Xunta de Galicia and the European Union (European Regional Development Fund- Galicia 2014-2020 Program), by grant ED431G 2019/01, and by Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), ERDF 2014-2020: Call ED431G 2019/04. DV is supported by a 2020 Leonardo Grant for Researchers and Cultural Creators from the BBVA Foundation. MG is supported by a Ramón y Cajal grant (RYC2019-028473-I).

to mislabelings in the gold-standard, as *Kiko* in “Camiño Neocatecumenal de *Kiko Argüelles*” (*sic*), annotated as I-PER instead of B-PER.

References

- Agerri, R., X. Gómez Guinovart, G. Rigau, and M. A. Solla Portela. 2018. Developing new linguistic resources and tools for the Galician language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).
- Agerri, R., I. San Vicente, J. A. Campos, A. Barrena, X. Saralegi, A. Soroa, and E. Agirre. 2020. Give your text representation models some love: the case for Basque. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4781–4788, Marseille, France, May. European Language Resources Association.
- Bengio, Y., R. Ducharme, P. Vincent, and C. Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.
- Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Cañete, J., G. Chaperon, R. Fuentes, and J. Pérez. 2020. Spanish pre-trained BERT model and evaluation data. In *Practical ML for Developing Countries Workshop (PML4DC) at ICLR*. Learning under limited/low resource scenarios.
- Collobert, R. and J. Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167.
- Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. 2011. Natural Language Processing (Almost) From Scratch. *Journal of Machine Learning Research*, 12:2493–2537.
- Dai, A. M. and Q. V. Le. 2015. Semi-supervised sequence learning. In *Advances in neural information processing systems*, pages 3079–3087.
- Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.
- Ettinger, A. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.
- Freixeiro Mato, X. R. 2003. *Gramática da Língua Galega IV. Gramática do texto*. A Nosa Terra, Vigo.
- Garcia, M. and P. Gamallo. 2010. Análise Morfosintáctica para Português Europeu e Galego: Problemas, Soluções e Avaliação. *Linguamática*, 2(2):59–67.
- Garcia, M., C. Gómez-Rodríguez, and M. A. Alonso. 2016. Creación de un treebank de dependencias universales mediante recursos existentes para lenguas próximas: el caso del gallego. *Procesamiento del Lenguaje Natural*, 57:33–40.
- Garcia, M., C. Gómez-Rodríguez, and M. A. Alonso. 2018. New treebank or repurposed? on the feasibility of cross-lingual parsing of romance languages with universal dependencies. *Natural Language Engineering*, 24(1):91–122.
- Guinovart, X. G. and S. L. Fernández. 2009. Anotación morfosintáctica do Corpus Técnico do Galego. *Linguamática*, 1(1):61–70.
- Guinovart, X. 2017. Recursos integrados da lingua galega para a investigación lingüística. *Gallaecia. Estudos de lingüística portuguesa e galega*, pages 1045–1056.
- IGE. 2018. Coñecemento e uso do galego. Instituto Galego de Estatística, http://www.ige.eu/web/mostrar_actividade_estadistica.jsp?idioma=gl&codigo=0206004.
- Jiang, N. and M.-C. de Marneffe. 2019. Evaluating BERT for natural language inference: A case study on the Commitment-Bank. In *Proceedings of the 2019 Con-*

- ference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6086–6091, Hong Kong, China, November. Association for Computational Linguistics.
- Karthikeyan, K., Z. Wang, S. Mayhew, and D. Roth. 2020. Cross-Lingual Ability of Multilingual BERT: An Empirical Study. In *International Conference on Learning Representations (ICLR 2020)*.
- Kingma, D. P. and J. Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR 2015)*. arXiv preprint arXiv:1412.6980.
- Kitaev, N. and D. Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia, July. Association for Computational Linguistics.
- Koutsikakis, J., I. Chalkidis, P. Malakasiotis, and I. Androutsopoulos. 2020. GREEK-BERT: The Greeks visiting Sesame Street. In *11th Hellenic Conference on Artificial Intelligence*. ACM.
- Kuratov, Y. and M. Arkhipov. 2019. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. *Computational Linguistics and Intellectual Technologies*, 18:333–339.
- Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *Proceedings of the International Conference of Learning Representations (ICLR 2020)*.
- Landauer, T. K. and S. T. Dumais. 1997. A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.
- Lee, S., H. Jang, Y. Baik, S. Park, and H. Shin. 2020. KR-BERT: A Small-Scale Korean-Specific Language Model. *arXiv preprint arXiv:2008.03979*.
- Lin, Y., Y. C. Tan, and R. Frank. 2019. Open sesame: Getting inside BERT’s linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy, August. Association for Computational Linguistics.
- Lindley Cintra, L. F. and C. Cunha. 1984. *Nova Gramática do Português Contemporâneo*. Livraria Sá da Costa, Lisbon.
- Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.
- Malvar, P., J. R. Pichel, Ó. Senra, P. Gamallo, and A. García. 2010. Vencendo a escassez de recursos computacionais. carvalho: Tradutor automático estatístico inglês-galego a partir do corpus paralelo europarl inglês-português. *Linguamática*, 2(2):31–38.
- McDonald, S. and M. Ramscar. 2001. Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 23.
- Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013a. Efficient estimation of word representations in vector space. In *Workshop Proceedings of the International Conference on Learning Representations (ICLR) 2013*. arXiv preprint arXiv:1301.3781.
- Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Nivre, J., M.-C. de Marneffe, F. Ginter, J. Hajič, C. D. Manning, S. Pyysalo, S. Schuster, F. Tyers, and D. Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France, May. European Language Resources Association.

- Ortiz Suárez, P. J., L. Romary, and B. Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online, July. Association for Computational Linguistics.
- Padró, L. 2011. Analizadores Multilingües en FreeLing. *Linguamatica*, 3(2):13–20.
- Pennington, J., R. Socher, and C. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.
- Peters, M., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.
- Pires, T., E. Schlinger, and D. Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July. Association for Computational Linguistics.
- Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.
- Rojo, G., M. López Martínez, E. Domínguez Noya, and F. Barcala. 2019. Corpus de adestramento do Etiquetador/Lematizador do Galego Actual (XIADA), versión 2.7. *Centro Ramón Piñeiro para a investigación en humanidade*.
- Salant, S. and J. Berant. 2018. Contextualized word representations for reading comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 554–559, New Orleans, Louisiana, June. Association for Computational Linguistics.
- Samartim, R. 2012. Língua somos: A construção da ideia de língua e da identidade coletiva na galiza (pré-) constitucional. In *Novas achegas ao estudo da cultura galega II: enfoques socio-históricos e lingüístico-literarios*, pages 27–36.
- Sanh, V., L. Debut, J. Chaumond, and T. Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Proceedings of The 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing at NeurIPS2019*. arXiv preprint arXiv:1910.01108.
- Schnabel, T., I. Labutov, D. Mimno, and T. Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal, September. Association for Computational Linguistics.
- Souza, F., R. Nogueira, and R. Lotufo. 2019. Portuguese named entity recognition using BERT-CRF. arXiv preprint arXiv:1909.10649.
- Strzyz, M., D. Vilares, and C. Gómez-Rodríguez. 2019. Viable dependency parsing as sequence labeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 717–723, Minneapolis, Minnesota, June. Association for Computational Linguistics.
- TALG. 2016. CTG Corpus (Galician Technical Corpus). TALG Research Group. SLI resources, 1.0, ISLRN 437-045-879-366-6.
- TALG. 2018. SLI NERC Galician Gold CoNLL. TALG Research Group. SLI resources, 1.0, ISLRN 435-026-256-395-4.
- Teyssier, P. 1987. *História da Língua Portuguesa*. Livraria Sá da Costa, Lisbon, 3 edition.

- Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention Is All You Need. arXiv preprint arXiv:1706.03762.
- Vilares, D. and C. Gómez-Rodríguez. 2018. Transition-based parsing with lighter feed-forward networks. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 162–172, Brussels, Belgium, November. Association for Computational Linguistics.
- Vilares, D., M. Strzyz, A. Søgaard, and C. Gómez-Rodríguez. 2020. Parsing as pretraining. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9114–9121. AAAI Press.
- Virtanen, A., J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter, and S. Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. arXiv preprint arXiv:1912.07076.
- Vulić, I., E. M. Ponti, R. Litschko, G. Glavaš, and A. Korhonen. 2020. Probing pre-trained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online, November. Association for Computational Linguistics.
- Wenzek, G., M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and E. Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France, May. European Language Resources Association.
- Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. 2019. HuggingFace’s Transformers: State-of-the-art Natural Language Processing. arXiv preprint arXiv:1910.03771.
- Wu, S. and M. Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online, July. Association for Computational Linguistics.